**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #2**

January 13, 2014

**Written problems:** To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Tuesday, January 21, 2014. See the instructions for peer grading of the homework that are posted on the web pages.

*On this (as all homeworks) Stata / R code and unedited Stata / R output is* ***TOTALLY*** *unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*

***Unless explicitly told otherwise in the statement of the problem, in all problems requesting “statistical analyses” (either descriptive or inferential), you should present both***

* ***Methods: A brief sentence or paragraph describing the statistical methods you used. This should be using wording suitable for a scientific journal, though it might be a little more detailed. A reader should be able to reproduce your analysis. DO NOT PROVIDE Stata OR R CODE.***
* ***Inference: A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to “Reporting Associations” for details.***

This homework builds on the analyses performed in homework #1, As such, all questions relate to associations among death from any cause, serum low density lipoprotein (LDL) levels, age, and sex in a population of generally healthy elderly subjects in four U.S. communities. This homework uses the subset of information that was collected to examine MRI changes in the brain. The data can be found on the class web page (follow the link to Datasets) in the file labeled mri.txt. Documentation is in the file mri.pdf. See homework #1 for additional information.

1. Perform statistical analyses evaluating an association between serum LDL and 5 year all-cause mortality by comparing mean LDL values across groups defined by vital status at 5 years using a t test that presumes equal variances across groups. Depending upon the software you use, you may also need to generate descriptive statistics for the distribution of LDL within each group defined by 5 year mortality status. As this problem is directed toward illustrating correspondences between the t test and linear regression, you do not need to provide full statistical inference for this problem. Instead, just answer the following questions.
   1. What are the sample size, sample mean and sample standard deviation of LDL values among subjects who survived at least 5 years? What are the sample size, sample mean and sample standard deviation of LDL values among subjects who died within 5 years? Are the sample means similar in magnitude? Are the sample standard deviations similar?

**Among subjects who survived five years, there were 606 observations, with a mean serum LDL of 127.198 mg/dl and a standard deviation of 32.929 mg/dl. There were 119 subjects who died within five years with a mean serum LDL of 118.698 mg/dl and a standard deviation of 36.157 mg/dl. The means differ by around 8.5 mg/dl which is not very much relative to the mayo-clinic LDL levels and the standard deviation differs by about 3.2, again, not a very large difference.**

* 1. What are the point estimate, the estimated standard error of that point estimate, and the 95% confidence interval for the true mean LDL in a population of similar subjects who would survive at least 5 years? What are the corresponding estimates and CI for the true mean LDL in a population of similar subjects who would die within 5 years? Are the point estimates similar in magnitude? Are the standard errors similar in magnitude? Explain any differences in your answer about the estimates and estimated SEs compared to your answer about the sample means and sample standard deviations.

**The point estimate of the true mean serum for subjects who survived at least five years is LDL is 127.198 mg/dl and the standard error of the point estimate is 1.338 mg/dl with a 95% confidence interval of 124.571-129.825vmg/dl. For subjects who died within five years, the point estimate of true mean serum LDL is 118.698 mg/dl with standard error of 3.315 mg/dl and a 95% confidence interval of 112.133-125.261mg/dl. The means, being identical to those in the previous question, are similar in magnitude; however, the standard errors are quite different from each other in magnitude with a difference of 2 but at this small size, the standard error for those who died within 5 years is nearly 2.5 that of the standard error for survivors. The point estimates of the true means are the same as the sample means because sample means are unbiased and their expected value is the true mean. The standard error is the standard deviation of the sampling distribution, this is produced by dividing the standard deviation of the sample mean by the square root of the sample size. This is the reason for the difference between the standard deviation and the standard error.**

* 1. Does the CI for the mean LDL in a population surviving 5 years overlap with the CI for mean LDL in a population dying with 5 years? What conclusions can you reach from this observation about the statistical significance of an estimated difference in the estimated means at a 0.05 level of significance?

**The CI for the mean LDL in the surviving subjects and the CI for mean LDL in the subjects who died within 5 years does overlap. This does speak to the statistical significance of an estimated difference in the estimated means at the .05 level of significance. If a confidence interval of one group contained the point estimate of the other group then it would mean a non-significant result and no overlap in the confidence intervals would indicate a significant result, but the overlap between confidence intervals does not indicate any conclusive result.**

* 1. If we presume that the variances are equal in the two populations, but we want to allow for the possibility that the means might be different, what is the best estimate for the standard deviation of LDL measurements in each group? (That is, how should we combine the two estimated sample standard deviations?)

**There is a commonly used estimate for the standard error of the difference in two group means when it is presumed that the variances are equal. It is done by taking the sum of the standard deviation of each group multiplied by the corresponding group’s sample size minus 1 and dividing that by the entire sample size minus two and taking the square root of the result, this is then multiplied by the square root of the sum of the inverses of the sample size of each group. For clarity, this would look like the following equation:**

**square-root[((n1-1)(sd1)+(n2-1)(sd2))/( n1+ n2)]\*square-root[(1/ n1+1/ n2)]**

**This is called a pooled variance estimate.**

* 1. What are the point estimate, the estimated standard error of the point estimate, the 95% confidence interval for the true difference in means between a population that survives at least 5 years and a population that dies with 5 years? What is the P value testing the hypothesis that the two populations have the same mean LDL? What conclusions do you reach about a statistically significant association between serum LDL and 5 year all cause mortality?

**The point estimate of the difference in mean serum LDL between subjects who survived 5 years or beyond and subjects who died within five years is -8.501 mg/dl with an estimated standard error of 3.567 mg/dl and a 95% confidence-interval of 1.911-15.090 mg/dl. The two-sided p-value for a difference in mean LDL is 0.012. Based on the two-sided p-value, we can conclude that at the p=.05 significance threshold, there is evidence to reject the null hypothesis that the two means are the same.**

1. Perform statistical analyses evaluating an association between serum LDL and 5 year all-cause mortality by comparing mean LDL values across groups defined by vital status at 5 years using ordinary least squares regression that presumes homoscedasticity. As this problem is directed toward illustrating correspondences between the t test and linear regression, you do not need to provide full statistical inference for this problem. Instead, just answer the following questions.
   1. Fit two separate regression analyses. In both cases, use serum LDL as the response variable. Then, in model A, use as your predictor an indicator that the subject died within 5 years. In model B, use as your predictor an indicator that the subject survived at least 5 years. For each of these models, tell whether the model you fit is saturated? Explain your answer.

**These models are both saturated. There are as many parameters in the regression model as there are groups between survivors and non-survivors. Because of this, the intercept of the model for survivors is the same as the estimate of the mean for survivors in the t-test. The same is true for those who died within 5 years.**

* 1. Using the regression parameter estimates from one of your models (tell which one you use), what is the estimate of the true mean LDL among a population of subjects who survive at least 5 years? How does this compare to the corresponding estimate from problem 1?

**Using the intercept of the model with the indicator variable for subjects who died within five years, my estimate for the true mean LDL of subjects who survive at least 5 years is 127.198 mg/dl which is the same as my estimate for the mean from problem 1.**

* 1. Using the regression parameter estimates from one of your models (tell which one you use), what is a confidence interval for the true mean LDL among a population of subjects who survive at least 5 years? How does this compare to the corresponding estimate from problem 1? Explain the source of any differences.

**Using the 95% confidence interval around the intercept of the model where the indicator is for subjects who died within five years, the confidence interval for mean LDL among subjects who survived at least 5 years is 124.528-129.868 mg/dl. The confidence interval uses pooled SE as described in question 1d rather than the SE for 5-year survivors which explains the minor differences between the two confidence intervals.**

* 1. Using the regression parameter estimates from one of your models (tell which one you use), what is the estimate of the true mean LDL among a population of subjects who die within 5 years? How does this compare to the corresponding estimate from problem 1?

**Using the intercept of the model with the indicator variable for subjects who survived at least five years, my estimate for the true mean LDL of subjects who died within 5 years is 118.698 mg/dl which is the same as my estimate for the mean from problem 1.**

* 1. Using the regression parameter estimates from one of your models (tell which one you use), what is a confidence interval for the true mean LDL among a population of subjects who die within 5 years? How does this compare to the corresponding estimate from problem 1? Explain the source of any differences.

**Using the 95% confidence interval around the intercept of the model where the indicator is for subjects who survived at least 5 years, the confidence interval for mean LDL among subjects who died within 5 years is 112.673-124.722 mg/dl. The confidence interval uses pooled SE as described in question 1d rather than the SE for 5-year survivors which explains the minor differences between the two confidence intervals.**

* 1. If we presume the variances are equal in the two populations, what is the regression based estimate of the standard deviation within each group for each model? How does this compare to the corresponding estimate from problem 1?

**The regression based estimate of the standard deviation within each group is the standard error for the regression coefficient. In both models A and B, these standard errors have the same value as the standard error of the difference in means from the t-test. This is an intuitive relationship because the regression coefficient estimates the difference in means between the two groups so the standard error of the regression coefficient is the same as the standard error for the difference in means.**

* 1. How do models A and B relate to each other?

**Since the models are saturated, the intercept of model A is the intercept of model B plus the regression coefficient of model B. Model B has a parallel relationship to model A.**

* 1. Provide an interpretation of the intercept from the regression model A.

**In model A, the intercept is the mean serum LDL for patients who have an indicator for death within 5 years of zero; in other words, the group who survived at least 5 years.**

* 1. Provide an interpretation of the slope from the regression model A.

**In model A, the slope is the difference in mean serum LDL between the group of subjects who survived at least 5 years and those that died within 5 years.**

* 1. Using the regression parameter estimates, what are the point estimate, the estimated standard error of the point estimate, the 95% confidence interval for the true difference in means between a population that survives at least 5 years and a population that dies within 5 years? What is the P value testing the hypothesis that the two populations have the same mean LDL? What conclusions do you reach about a statistically significant association between serum LDL and 5 year all cause mortality? How does this compare to the corresponding inference from problem 1?

**The point estimate of the difference in mean serum LDL between subjects who survived 5 years or beyond and subjects who died within five years is -8.501 mg/dl with an estimated standard error of 3.567 mg/dl and a 95% confidence-interval of 1.911-15.090 mg/dl. The two-sided p-value for a difference in mean LDL is 0.012. Based on the two-sided p-value, we can conclude that at the p=.05 significance threshold, there is evidence to reject the null hypothesis that the two means are the same. This is exactly the same as the inference for the t-test based analysis.**

1. Perform statistical analyses evaluating an association between serum LDL and 5 year all-cause mortality by comparing mean LDL values across groups defined by vital status at 5 years using a t test that allows for the possibility of unequal variances across groups. How do the results of this analysis differ from those in problem 1? (Again, we do not need a formal report of the inference.)

**Using a t-test with a standard error of the difference of means that allows for unequal variances, I get an estimated mean serum LDL for subjects who survived at least 5 years of 127.198 mg/dl, standard error of 1.338 and CI of 124.571-129.825 mg/dl and mean serum LDL for subjects who died within 5 years of 118.698 mg/dl, standard error 3.315 and CI of 112.133-125.261mg/dl. My difference of means is estimated as 8.501 with a standard error of 3.574 and a confidence interval of 1.44-15.560. The two-sided p-value is 0.019. The standard error for the difference is a little larger and so is the range of the 95% confidence interval because of the different method for estimating the standard error. The p-value is a little higher, but not so much that it changes our previous conclusion that at alpha=.05 we have evidence to reject the null hypothesis of no statistically significant difference of mean LDL between groups.**

1. Perform statistical analyses evaluating an association between serum LDL and 5 year all-cause mortality by comparing mean LDL values across groups defined by vital status at 5 years using a linear regression model that allows for the possibility of unequal variances across groups. How do the results of this analysis differ from those in problem 3? (Again, we do not need a formal report of the infrence.)

**Using the indicator variable for 5-year surival, I have a regression coefficient corresponding to the difference in mean serum LDL between survivors and subjects who died within 5 years of 8.501 mg/dl with a robust standard error of 3.566 mg/dl and a 95% confidence interval of 1.500-15.501 mg/dl and a p-value testing against having a regression coefficient of 0 of 0.017. The difference in standard error, p-value and confidence interval is due to the fact that the regression based robust standard error is estimated using Huber-White Sandwich estimator as opposed to the pooled standard error.**

1. Perform a regression analysis evaluating an association between serum LDL and age by comparing the distribution of LDL across groups defined by age as a continuous variable. (Provide formal inference where asked to.)
   1. Provide descriptive statistics appropriate to the question of an association between LDL and age. Include descriptive statistics that would help evaluate whether any such association might be confounded or modified by sex. (But we do not consider sex in the later parts of this problem.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Serum Low Density Lipoprotein (LDL) | | | |
|  | < 129 mg/dL | 130 - 159 mg/dL | > 160 mg/dL | Any Level |
|  | (n=393) | (n=225) | (n=107) | (n=725) |
| Male (%) | 56% | 43% | 42% | 50% |
| Age\* | 74.697 (5.252; 65-92) | 74.196 (5.624; 67-99) | 74.879 (5.77; 65-94) | 74.568 (5.446; 65-99) |
| \*Statistics reported are mean (Standard Deviation; Minimum-Maximum) | | | |  |

**As seen in the table above, there are again 10 missing values for ldl which are excluded from the table and the analysis. There are more subjects in the lower levels of LDL and more males in the lower levels of LDL, which might mean that sex is a confounder or effect modifier. Without regard for sex, the distribution of age is pretty consistent across categorical LDL levels.**

* 1. Provide a description of the statistical methods for the model you fit to address the question of an association between LDL and age.

**To evaluate an association between LDL and age, I am going to fit a linear regression that compares mean ldl across age. Since I want to be anti-conservative in my inference, I am going to use robust standard errors using the Huber-White Sandwich Estimator. Age will be the predictor of interest (POI) and the outcome will be mean serum LDL.**

* 1. Is this a saturated model? Explain your answer.

**This model is not saturated, as there are many more values for my predictor of interest than there are parameters.**

* 1. Based on your regression model, what is the estimated mean LDL level among a population of 70 year old subjects?

**Mean serum LDL for 70 year olds is 126.215 mg/dl**

* 1. Based on your regression model, what is the estimated mean LDL level among a population of 71 year old subjects? How does the difference between your answer to this problem and your answer to part c relate to the slope?

**Mean serum LDL for 71 year olds is 126.125 mg/dl. This difference compared to 70 year olds mean LDL of -0.090 mg/dl is 1\*the slope of the regression.**

* 1. Based on your regression model, what is the estimated mean LDL level among a population of 75 year old subjects? How does the difference between your answer to this problem and your answer to part c relate to the slope?

**Mean serum LDL for 75 year olds is 125.764 mg/dl. This difference compared to 70 year olds mean LDL of -0.451 mg/dl is 5\*the slope of the regression.**

* 1. What is the interpretation of the “root mean squared error” in your regression model?

**This is the standard deviation of the difference between our predicted LDL and the observed LDL across age groups.**

* 1. What is the interpretation of the intercept? Does it have a relevant scientific interpretation?

**The intercept is the extrapolated mean LDL for people aged 0. This does not have a relevant scientific interpretation because there were no newborns in our dataset and the fit is very bad.**

* 1. What is the interpretation of the slope?

**The slope is the difference in mean serum LDL corresponding to a 1 year positive difference in age.**

* 1. Provide full statistical inference about an association between serum LDL and age based on your regression model.

**Method: I used a linear regression with robust standard errors using the Huber-White sandwich estimators with predictor age and outcome mean serum LDL.**

**Inference: My regression estimated a linear relationship between age and mean serum LDL of -.090 mg/dl, this had a standard error of .233 mg/dl, a 95% confidence interval of -.547-.367 mg/dl, and a p-value testing a null hypothesis of no linear relationship of .698. Our results indicate that we cannot at the p=.05 level of significance (or any other meaningful level!) reject the null hypothesis that there is no linear relationship between age and mean serum LDL.**

* 1. Suppose we wanted an estimate and CI for the difference in mean LDL across groups that differ by 5 years in age. What would you report?

**Method: By multiplying the regression coefficient and standard error by 5 created a new point estimate of the regression coefficient for a difference of five years of age and a confidence interval based on point estimate plus or minus 1.96\*standard error.**

**From my simple modifications I produced a regression slope of -.451 mg/dl with a standard error of 1.147 mg/dl and a 95% confidence interval of -2.70 to 1.80 mg/dl. Because the confidence interval crosses zero, I would report that I did not find evidence to reject the null hypothesis of no linear relationship between mean serum ldl and age in five year increments.**

* 1. Perform a test for a nonzero correlation between LDL and age. How does your regression-based conclusion about an association between LDL and age compare to inference about correlation?

**I got a correlation of -.0146 with a p-value, using the Pearson test, of 0.694. This is different in magnitude but ultimately gives the same result that we do not see a significant (alpha=.05) linear relationship between ldl and age.**

**Discussion Sections: January 13 – 17, 2014**

We will discuss the dataset regarding FEV and smoking in children. Come do discussion section prepared to describe the approach to the scientific question posed in the documentation file fev.doc.